SOME WELL-KNOWN MENTAL TESTS EVALUATED AND COMPARED

BY
DOROTHY RUTH MORGENTHAU

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, in the Faculty of Philosophy, Columbia University

REPRINTED FROM
ARCHIVES OF PSYCHOLOGY
R. S. WOODWORTH, EDITOR

No. 52

NEW YORK
MAY, 1922

SOME WELL-KNOWN MENTAL TESTS EVALUATED AND COMPARED

BY
DOROTHY RUTH MORGENTHAU

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, in the Faculty of Philosophy, Columbia University

REPRINTED FROM
ARCHIVES OF PSYCHOLOGY
No. 52

NEW YORK
MAY, 1922

TABLE OF CONTENTS

ACKNOWLEDGMENT.

For the advice of Professor Edward Lee Thorndike of Teachers College, Columbia University, and of Dr. William Healy and Dr. Augusta Bronner of Judge Baker Foundation, Boston, the writer wishes to express appreciation. Special thanks are due for the painstaking assistance given by Professor Robert S. Woodworth of Columbia University, New York City.

Some Well-Known Mental Tests Evaluated and Compared

ONE who approaches the subject of the measuring of children's mentality will find that the mind of the normal child has received attention in what we may call vertical and parallel respects. There have been a considerable number of tests developed by students of psychology in the endeavor to secure mental measurements independent of the experience and judgment of the clinician. The development has been in a vertical manner, that is, the best recognized psychologists who have undertaken this work have each developed tests, have each put them into extensive practice and have published the results of that experience. But each of these psychologists has developed his test on his own suppositions, and, basing the nature of his test on his own experience, has tried to evolve a plan of testing which is supposed to be useful in determining mental conditions of such general extent that they may roughly be called intelligence. Thus we have the Stanford-Binet scale, the most generally used of any one of the mental tests. Then there are the Porteus tests, the Pintner-Patterson performance scale, and a dozen or more of others which are known to every clinical psychologist.

The development of mental tests has been parallel in that none of these psychologists in developing their own ideas have carried them to the point of thoroughly comparing the results obtained by their tests to the results obtained by the simultaneous use of a number of the other tests all with respect to normal children. There has been some comparison of results of the various tests when applied to abnormal children but this has not been thoroughgoing and has been done not by making the tests with the idea of eventually combining the results and of placing valuations upon them, but merely in the course of clinical work with abnormal children. It is questionable whether such results are sufficiently thorough to be considered the basis for a convincing answer as to the relative value of the respective tests, and inasmuch as they were made on abnormal minds, one would not dare to trust even those

comparative results with respect to what the test will show as to normal minds.

Those who have developed their respective tests have compared them with *some* other mental test, most frequently one of the Binet revisions. But, no considerable number of the tests which have been so developed in parallel fashion have been applied purposely to obtain comparative results and to ascertain which if any of them can be shown to be untrustworthy and what group of them can be relied upon as furnishing a satisfactory schedule for testing and comparing the common elements of mentality in normal children.

Upon perceiving that there was a lack of any purposely made comparative study of mental tests it was proposed herein to set forth the results of such a study of about a dozen of the most commonly used mental tests. The tests were applied to a large number of unselected normal children, in general each child receiving the full schedule of tests. By means of the results to be obtained from this comparative study it was anticipated:

1. That the degree of reliability of each test would be indicated.
2. That the same purposes could be effected with respect to the value of each test.
3. That the information obtained under the first and second headings would make it possible to select a schedule of tests of indicated reliability for application to normal minds, or further, whether the Stanford-Binet alone would suffice.
4. That by restricting the ages of children tested in general to from ten to sixteen years, the period in which individual capacities first assume importance for vocational determination, it would be possible to guide the vocational training with some degree of success.

A brief statement of the results can now be given reserving the more detailed statement involving the basis and methods for the results for future pages. The first aim, to secure an estimate of the reliability of the tests used, was largely successful. Of the thirteen tests, the reliability of which was investigated, one class, the four construction tests, Healy A and B. and Knox Moron tests, and diamond shaped frame, were found to be unreliable; five other tests were found to be reliable, namely the Stanford-Binet, Pintner Non-Language group test, Thorndike Reading Scale Alpha 2, Porteus Maze test, and

Tapping test; while the reliability of four tests, the Myers Mental Measure, Healy Pictorial Completion test II, Healy-Bronner learning tests, and the Crossline test was undetermined for various reasons.

The results obtained as to value of the tests were as follows:

Stanford-Binet, Pintner, Alpha II, and Porteus are valuable tests and should be included in individual case studies. In spite of their unmeasured reliability, Myers and Pictorial Completion II are also valuable tests and should likewise be included. Judgment should be suspended with regard to learning tests. The Tapping test is of doubtful value and its use should be left to the discretion of the examiner. The Construction tests because of their unreliable character do not give valuable results.

As to the schedule of tests to be used in testing normal minds it was found best not to use the Stanford-Binet alone but to have the schedule composed of that test and the five others which were found valuable. From the tests used and results obtained it cannot be stated here whether this schedule is of value as to vocational guidance for the reason that the factors involved in each test are not known with certainty and until they are known, definite valid conclusions about the abilities of the individuals concerned cannot be reached.

SUBJECTS.

It was desired to test one hundred normal but otherwise unselected children. In order to obtain an unselected group it proved necessary to select the subjects very carefully, for, if all the children tested had been from a Children's Home, or from a Settlement, or from any one school, the result would have been a highly selected group. To avoid this a few were taken from many different sources and in this respect the distribution proved to be reasonably satisfactory.

As to age, originally the plan was to have about ten children at each of the ten periods of one year each, from seven to sixteen inclusive. But this plan was given up because our interest is not with the six or seven year old who has to go to school and learn fundamentals, no matter wherein his is gifted and who rarely shows talents or handicaps at such an early age. Our chief concern is with children in the sixth, seventh or eighth grades and in high school, because they are the adjustment problems, and because it is important to aid them if possible in deciding whether they should remain in school or go to work. If the latter what should they do, if the former what sort of training do they need? So the attempt was made to lay all the emphasis here and reduce the number of children under eleven to a minimum. Another objection to the original plan is that ten in a group is too small for any kind of generalization.

The total number tested was 128, of which 116 usable records were retained. For various reasons many of these records are incomplete so that this number was necessary in order to have a minimum of 100 scores on each test. There still remain some tests which were given to less than 100 children, but the number is in each case sufficiently large to give valuable results.

All defectives were excluded, for in mixing their records with those of normal children many confusions would have arisen, and the issues would have been less clear. Much intensive work has been done in testing defectives, so that we know a great deal about their reactions to a group of tests such as we have chosen. To be sure, they vary considerably in their results, but we know in general the points where they are weakest as in abstract reasoning and formal generalization,

and also the points in which proportionately they excel. By narrowing the field to normals the significance of the conclusions can be made more pertinent. This was an arbitrary procedure dependent largely upon the judgment of the writer, and subject to criticism on this basis. It is quite possible that some very dull normals were also excluded, this being justified on the grounds that their normality might reasonably have been called in question by more severe examiners. With reference to the three cases whose I. Q.'s fall below 80, there seems to be no doubt that they are to be considered as dull normals. The grade they attained in school for their age, their response on the other tests and their behavior in the community all argue for including them in our study. The boy receiving the lowest I. Q.—73—was born in the United States but taken to Italy at the age of five, and remained there six years. In spite of this he was in the eighth grade. He did very well with all the construction tests.

As no limitations were set at the other end, the grade and I. Q. distributions are higher than one would otherwise expect in a general sampling of the population.

I. Thirty-seven children, twelve girls and twenty-five boys, were tested at the Home for Jewish Children in Dorchester, Massachusetts. Many of these children were half orphans, some had lost both parents—most of them were in the Home temporarily. They were chosen from the total number entirely by chance. They all attended public school in the vicinity and all but two or three had come to the Home within two years. All were able to speak and understand English, this being the only language used at the institution, although in many of their homes no English was spoken. Their ages ranged from 7-0 to 15-1.

II. Twenty-four girls came from Frances Willard Settlement in Boston, Massachusetts. These were divided into three clubs—one consisting of one seventh grade and ten eighth grade girls, the youngest being 12-7 and the oldest 14-2. They came one evening a week for the express purpose of taking the tests. They were the first ones to volunteer from a large group. The other two groups of seven and six respectively were younger girls who happened to meet on afternoons which were convenient for the examiner.

III. Six high school girls in New York volunteered to take the tests.

IV. The ninth grade consisting of six boys and five girls in the Woodmere School (private) at Woodmere, Long Island, were tested. The ages ranged from 13-1 to 15-2.

V. The poorer section of the 8B class of Public School 11, New York, were tested. There were thirty boys in the class ranging in age from 13-2 to 16-11.

VI. Finally eight miscellaneous children were tested.

The subjects selected appeared to give a satisfactory difference in quality so as to bring out the capacities of the tests to meet a variety of normal mental conditions.

TABLE I AGE DISTRIBUTION 116 CASES

Yrs.	Mos.	Frequency		Yrs.	Mos.	Frequency		Yrs.	Mos.	Frequency	
7	0	1		10	4	2		13	8	0	
7	1	0		10	5	1		13	9	2	
7	2	0		10	6	0		13	10	3	
7	3	0		10	7	0		13	11	5	19
7	4	0		10	8	0		14	0	1	
7	5	0		10	9	0		14	1	1	
7	6	1		10	10	0		14	2	2	
7	7	0		10	11	1	6	14	3	1	
7	8	0						14	4	2	
7	9	0		11	0	1		14	5	2	
7	10	1		11	1	0		14	6	2	
7	11	0	3	11	2	0		14	7	3	
8	0	0		11	3	2		14	8	0	
8	1	0		11	4	1		14	9	0	
8	2	0		11	5	1		14	10	1	
8	3	0		11	6	2		14	11	6	21
8	4	0		11	7	0		15	0	1	
8	5	0		11	8	0		15	1	2	
8	6	0		11	9	0		15	2	5	
8	7	0		11	10	1		15	3	1	
8	8	0		11	11	2	10	15	4	0	
8	9	0						15	5	2	
8	10	1		12	0	3		15	6	2	
8	11	1	2	12	1	1		15	7	2	
9	0	1		12	2	1		15	8	0	
9	1	0		12	3	2		15	9	0	
9	2	1		12	4	0		15	10	0	
9	3	0		12	5	2		15	11	1	16
9	4	1		12	6	0		16	0	3	
9	5	0		12	7	2		16	1	0	
9	6	0		12	8	1		16	2	1	
9	7	2		12	9	2		16	3	1	
9	8	1		12	10	4		16	4	1	
9	9	1		12	11	3	21	16	5	0	
9	10	1		13	0	1		16	6	1	
9	11	0	8	13	1	2		16	7	1	
10	0	0		13	2	1		16	8	1	
10	1	0		13	3	0		16	9	0	
10	2	0		13	4	0		16	10	0	
10	3	2		13	5	3		16	11	1	10
				13	6	1					
				13	7	1					
		15				44				57	

Distribution of the subjects by age.

It will be noted that only 19 of the 116 subjects are under eleven years old.

TABLE II

GRADE DISTRIBUTION—114 CASES.

2 had left school.

Grade	Frequency
I	1
II	2
III	2
IV	11
V	7
VI	16
VII	10
VIII	44
IX or I H. S.	12
X or II H. S.	4
XI or III H. S.	5
XII or IV H. S.	0
Left School	
VIII	1
II H. S.	1

The vast majority of subjects were in the VIth to IXth grades inclusive.

INTELLIGENCE QUOTIENT DISTRIBUTION. 112 CASES.

Scale:—1 square to 1 child
70 means 70.000 to 79.999 etc.
The curve of distribution is skewed positively.

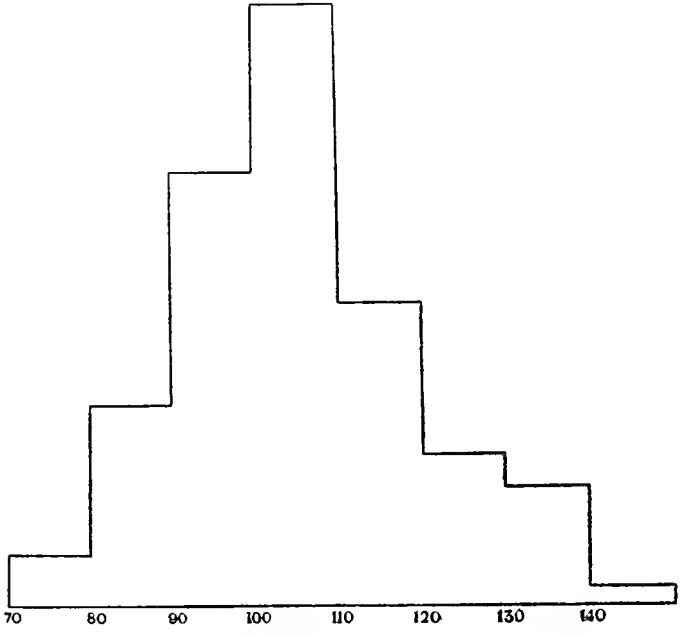


TABLE III

DISTRIBUTION OF INTELLIGENCE QUOTIENTS—112 CASES.

I. Q.	Frequency	I. Q.	Frequency	I. Q.	Frequency
70	0	95	2	120	0
71	0	96	3	121	1
72	0	97	4	122	2
73	1	98	2	123	0
74	1	99	3	124	0
75	0	100	7	125	1
76	1	101	7	126	2
77	0	102	4	127	2
78	0	103	2	128	1
79	0	104	0	129	0
80	1	105	5	130	3
81	0	106	4	131	1
82	0	107	3	132	1
83	0	108	2	133	0
84	1	109	2	134	1
85	2	110	1	135	0
86	2	111	1	136	1
87	1	112	1	137	0
88	2	113	3	138	0
89	3	114	2	139	0
90	1	115	1	140	0
91	2	116	4	141	1
92	1	117	1	142	0
93	4	118	2	143	0
94	4	119	2	144	0

4 were not given the Stanford-Binet test.
Average 104.5
Mental age in months
Average 154.8
Mean Square Deviation 34.54

The table shows that very few of the children tested had I. Q.'s below normal.

Age-Grade Distribution

CHRONOLOGICAL AGE

	6.5	7.5	8.5	9.5	10.5	11.5	12.5	13.5	14.5	15.5	16.5	Total cases
I		1										1
II		2										2
III			1									1
IV			1	6	3	2						12
V				2	3	1	1					7
VI						6	10					16
VII						1	5	2	2			10
VIII							5	9	15	11	6	46
IX								7	3	1		11
X								1	2	2		5
XI										1	4	5
	Total	3	2	8	6	10	21	19	22	15	10	116

CHRONOLOGICAL AGE

Chronological age—mental age distribution.

	6.5	7.5	8.5	9.5	10.5	11.5	12.5	13.5	14.5	15.5	16.5	Total cases
6.5												
7.5		1										1
8.5		2		1	1	1						5
9.5				3	1	1	1					6
10.5			1	2	3	2	2					10
11.5			1			1	3				1	6
12.5				1	1	3	4	2	3		1	15
13.5							4	1	4	1		10
14.5						2	3	2	6	6	3	22
15.5								2	4	8		14
16.5							3	4	2	1	1	11
17.5								2	1		1	4
18.5								3	1		1	5
19.5								1			2	3
Total		3	2	7	6	10	20	17	20	16	10	112

Mental age—grade distribution.

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	Total cases
6.5												
7.5	1											1
8.5		2		1	1	1						5
9.5				3	2	2						7
10.5			1	6	1	1	1					10
11.5				1	1	2	2	1				7
12.5					2	4	2	6				14
13.5						3	2	5				10
14.5						4		15	2		1	22
15.5								11	1	1		13
17.5								1	1	1	1	4
16.5							1	4	5	1		11
18.5								1	2		1	4
19.5									1		2	3
Total	1	2	1	11	7	17	8	44	12	3	5	3

TESTS BRIEFLY DESCRIBED AND REASONS FOR THEIR SELECTION

The large number of tests available had to be classified so as to find which tests covered identical ground; only one of these was then selected. Time was an element particularly to be recorded since preferably less than three hours should be devoted to each child for the completion of all tests. This allotment of time is considered by most authorities to be generous, particularly since the Stanford-Binet takes nearly three quarters of an hour, thus leaving only two hours for all the other tests. Consequently between alternate tests apparently serving the same purpose the briefer one was chosen. The same limitation on the amount of time to be spent on any one individual caused the necessary omission of some tests which were highly desirable except as to their length. In the last mentioned class are group tests requiring an hour or more to be applied. Where the results that are sought can be reached by group tests doubtless much time can be saved in using them, but the inquiries involved herein were such as to necessitate largely individual testing.

In selecting the tests another danger that was realized and that it was attempted to avoid, although, as the results show, not with entire success, was that a great many tests involved many sides of mental activity so that the final result expressed numerically would not be indicative of which mental abilities had tested favorably and which unfavorably. For instance, ability to deal with abstract and with concrete material may be extreme opposites giving a correlation of minus 100. If both kinds of material are combined in one test, the child who succeeds in one may fail in the other and vice versa. In computing the final scores compensation will give the same net result to two children of exactly opposite capabilities. If general intelligence is what we want we may find it in this way, but if we are interested in special abilities or disabilities these tests which hide them must not be used. We have found this confusion to exist in many tests, of course, never in such an extreme form as in the illustration above, and undoubtedly introduced on purpose, but we feel that its value is at all times questionable. This error is extremely difficult to eliminate

completely, in fact we can not be sure even now, as it will appear later in the results, that we have successfully done so.

Another source of error too often overlooked was borne in mind in selection of the tests, namely the variability of the test that is being considered. Where the same test is applied to a person at intervals and it is found that the resulting scores are not identical the question arises whether the varying scores can be combined so as to give a reliable standard for use and comparison with the results obtained when the test is applied to other children, or whether the variation indicates an unreliability in the test itself sufficiently serious to warrant the test being discarded. As an example of variations of such minor character that their existence does not indicate unreliability, and which can be compensated, we can take the tapping test where there may be a variation of about five taps in each direction from the average, which would be entirely satisfactory. Such variations are due to unessential and insignificant details of the conditions under which the test is repeated, such as posture of the child being tested, kind of pencil or stylus being used, etc. Taking ten or more measures of tapping ability would increase the reliability but the final results would show such slight difference from the result of one or two trials that the frequent repetition is entirely uncalled for to secure reasonable reliability.

On the other hand, if the variations in result obtained by repeated use of a test on the same individual are not of a minor character and if the day-to-day variability is so erratic that the variation is all the way from good performance to poor performance, then the situation is either that the child tested is shown to be subject to mental disturbance, or that the test itself shows a high and dangerous variability. If it is the test that is variable, it is obviously essential to weed it out *ab initio*. Such variability has been found to exist in the Knox cube test, in the application of which a uniformly normal child may make the record of an imbecile one day and of a super-normal child the next day. Of course, such a test, if not eliminated, would lead to results that are valueless for comparative purposes and dangerous for diagnostic ones.

As to variability, the reliability of a number of tests was established and recorded before the study was undertaken. As to the remaining tests, in order to overcome the possible existence of variations indicating unreliability it was necessary to

retest each child with the same or with a similar test after an interval of a week—no less or practice effect would be met, no more to avoid the effect of any mental growth in the interval.

The necessity of retesting caused by possible variability in the test itself, led to the subordinate but difficult problem of determining what methods of retesting would avoid errors due to the process itself. Thus, as has been mentioned, retesting must be done in such a manner as to avoid practice effect. It has been shown by various workers that certain types of tests once solved, such as most puzzles, are no longer tests at all, whereas others, such as auditory memory for digits and psychomotor control, show a minimum effect, which, after the week between tests, is negligible. Those of our tests which come within the last-mentioned class were similarly repeated. Those which were of the former type had similar tests substituted for them in the second trial, while still others falling between these classes were altered in details so that the same test could be repeated, avoiding the memory aspect.

The tests finally selected were:

1. *The Stanford revision of the Binet-Simon scale.* This test is so widely known that it does not seem to be necessary to describe it here.

2. *Pintner's mental survey non-language group test,* with Myers Mental Measure as an alternate for repeating. These tests involved a minimum use of language. In the Pintner test no language is used in the performance, and in fact it is possible to give this test to foreigners or deaf children through the medium of signs, while in giving the Myers Mental Measure it is necessary for the subject to understand simple language, but none is used in executing the test. The Pintner test has six parts, the first resembling the Knox cube test, the second and third being substitution tests, the fourth a drawing completion, while the fifth is a reversed drawing test, and the sixth a picture reconstruction. Following directions, Pictorial Completion, and two tests of picking out objects with common elements, compose the Myers test.

3. *Thorndike's reading scale Alpha 2.* This is a test in which language plays a prominent part. The subject reads a paragraph and then reads certain questions based upon the paragraph to which he writes his answer. To succeed he must understand the context of the paragraph, he must understand the question and know what it calls for, and he must be able

to find the answer in the context and write it down. This is a graded test which is applicable from the second grade through high school. Since the practical work of this research was undertaken, Dr. McCall of Teachers College has considerably increased the usefulness of this test by devising ten sets identical in method but with different contents, of which the test here used is one. It is now known as the Thorndike-McCall reading scale and its reliability has been thoroughly established.

4. *Healy's Pictorial Completion Test B* is an apperception test with the language element omitted. The ten pictures (plus one sample) present a day's activities of a young school boy, in which each picture contains a situation known to every child, such as eating breakfast, the school cloak room, a street accident, etc. In each picture one important element is lacking; pieces which complete the picture, plus fifty more of the same size being arranged in a definite order in a box from which the subject is at liberty to choose those which he desires. A clue to the missing piece is furnished by the pictures.

5. *Porteus Maze Tests.* Vineland Revision 1919. These tests are supposed to measure social fitness and common sense. Among the capacities which they were devised to measure are forethought and planning capacity, prudence and mental alertness in meeting a situation new to experience. There are eleven mazes, graded in difficulty from year three to fourteen. Beginning with year five, avoidance of blind alleys is the main requirement for a successful performance. The more complex the maze, the further ahead must one look in order to be certain that one is choosing the correct path. There is no time limit; in fact no mention of speed is made, and if the child asks he is told to do it as well as possible, taking as long as he likes. Porteus says that children fail mainly because of impulsiveness in action, overconfidence and carelessness, lack of pre-consideration, lack of planning capacity, irresolution and mental confusion, inability to sustain attention, or to profit by past mistakes.

6. *Tapping Tests—Healy's Form.* This consists of a sheet containing one hundred and fifty half inch squares, arranged ten in a row—fifteen rows. The subject taps once in each square, without touching the lines and covers as much ground as he can in thirty seconds. This is a simple test of psychomotor control which was repeated without alteration.

This test in a slightly different form was first introduced by Cattell in 1896, for testing freshmen at college. He had one hundred 1 cm. squares, into each of which the student must put a dot, completing the task as quickly as possible. Time was recorded; evidently there were no errors. This test was supposed to measure rate of movement. Clark Wissler used it with many of Cattell's other tests in his "Correlation of Mental and Physical Tests" on college freshmen in 1901. He found that the average time for men was 34 seconds, for women 30.8 seconds. In 1911 Whitley: (M. T. Whitley, An Empirical Study of Certain Tests for Individual Differences) reports results on Cattell's test, in which she kept the time constant (30 seconds) but computed the length of time which it would take to complete the blank. We have found the additional fifty squares useful in that some of our cases marked over one hundred squares in the thirty second time limit.

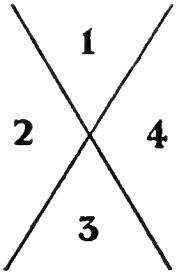
7. *Healy's Construction Tests A. and B.* The Knox-Moron test and Knox Modification of Healy A—a diamond-shaped frame, were used as alternates. We have called these A and B respectively to correspond with the Healy tests and for convenience. The equipment for these tests consists of a board containing one or more openings into which the child tested is supposed to fit pieces of wood so shaped that when properly arranged they will just close up the apertures. An advantage of these tests is the convenient size of the materials required. As all materials had to be carried from place to place the use of clumsy form boards or the tapping board with its dry batteries, metal plate and stylus, was practically out of the question. Where other things were equal, tests having the least paraphernalia were to be preferred.

8. *The Crossline Tests* shown in the figure were also given.

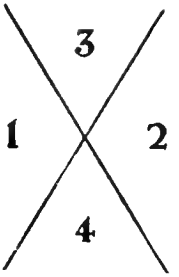
The crossline tests were included because they are a modification of the famous Code test, which is generally considered one of the best in the whole Stanford-Binet series. They take very little time to give and can easily be modified for repetition.

9. *Healy and Bronner Learning Tests.*—These tests were devised to test learning ability, not as in the skill experiment, but as it is found essential in the elementary school subjects. Learning test A—the association of two symbols, a figure and a number, resembles other substitution tests such as those of Woodworth and Wells, Pintner, and especially Woolley. The

I. Crossline Test



(a)



(b)

II. Crossline Test

1	4	7
2	5	8
3	6	9

(c)

1	2	3
4	5	6
7	8	9

(d)

(a) and (c) are the forms used generally.
(b) and (d) were used for retesting.

difference lies in the fact that three trials were given and speed of learning determined success. Learning test B is the association of a symbol with a sound, as in learning a language. The symbols are from the Phoenician alphabet, and the sounds consist of one or two consonants and a vowel, simple enough to pronounce but without meaning. This prevents older children from forming associations which would be impossible for those who did not know the meaning of the syllables. Test C is the association of a symbol and a value presented audibly, and test D is the association of ideas with a picture. The first three test a sort of rote ability whereas the latter tests learning of ideas.

It seems reasonable that success in school work may depend as largely upon learning ability as upon mental capacity, especially in the early grades where the chief requirement in most of our schools is a good rote memory, as in learning multiplication tables, and these two do not necessarily go together. Certain clinical cases bear out this suggestion, and these tests

were included to ascertain the reactions of normal unselected children in this respect.

National Intelligence tests were not yet published in November and December, 1919, when this study was begun, or they would surely have been considered and very likely used.

METHOD—APPLYING TESTS TO SUBJECTS

General Considerations

All of the tests except the non-language group tests and the Thorndike Alpha 2 were given to the subjects individually. The non-language tests were given sometimes individually and sometimes in groups of about ten with one exception where thirty eighth grade boys were tested in a group.

The time of day at which the tests were given varied considerably. About fifty of the subjects, from the Children's Home and from the Settlement, were tested in the evening. All others were tested in the daytime.

Care was taken to avoid giving any tests while the subject might be fatigued. Each child was questioned regarding the matter and whenever there were indications of fatigue the testing was always postponed.

Usually a subject was tested for only an hour and a half at one time; frequently the duration of the testing was shorter and only occasionally was it longer.

The tests were all scored according to the directions laid down by their respective authors. They were all scored personally by the examiner twice. In all of the tests selected for use the scoring is objective and requires no technique. Where possible, score cards or keys were used. Where the time taken by a subject to complete a test was to be recorded, the timing was done by means of a stop watch.

Much effort was expended in persuading the subjects to give an equal amount of attention and concentration to all of the tests, so that the results would not be affected by individual preferences. For a large proportion of the subjects the incentive of vocational guidance was offered and some general vocational advice, based partly on the experience of the examiner as well as on the tests, was given at the conclusion of the testing. Younger children needed no incentive and their enthusiasm was so pronounced that they continually applied to take more than the regular number of tests.

Supplementary information concerning the subjects was gathered and recorded, especially age in months, school grade, success in school work, marks, standing in class, whether a

repeater and how often, whether subject skipped any grades, etc. The vocational plans and interests of the older children were obtained whenever they had any. Results of physical examination were obtainable for a large per cent of the cases. Several subjects had also been given neurological examinations. Occasionally some result can be explained by reference to these findings, as for instance an unaccountably poor performance on the Healy Pictorial Completion test which was probably due to uncorrected vision. One case where peculiar results were obtained from the tests was explained by the physical examination which showed a history of epilepsy and thereupon the case was no longer considered.

Specific Observations on the Application of the Tests Selected

Stanford-Binet.—In the United States there have been several revisions of the Binet-Simon test, the most recent and well the best of these being that by Professor Lewis M. Terman of Leland Stanford University, California, published in its final form in 1916. This revision, called the Stanford-Binet, was the one used in this study. The score obtained in the Stanford-Binet test is expressed in years and months, mental age. This mental age, when divided by the life age, results in the intelligence quotient, which is expressed as a decimal. There have been some wrongful uses of the intelligence quotient. It is an attractive but erroneous idea that a certain intelligence quotient can be found below which all can be considered feeble minded while all above are normal or supernormal. The error in this idea has been pointed out by Fernald, Mateer, Kohs, and others, who demonstrate the degree of overlapping, and show how valueless the I. Q. is when reported without reference to life age.

The Stanford-Binet results can be analyzed, as well as summed up in the I. Q., and it is possible that a detailed analysis of the data would yield all the information required. The plea that general intelligence scales have a right to be so called is largely based upon the supposition that the functions which are tested are manifold. Auditory memory for rote material and for ideas, visual memory, language ability, reasoning ability, apperceptions, general information and many other abilities—all are found within the total range of tests. Unfor-

tunately, in the Stanford scale no child gets tested in all these fields, and further, since they are not standardized separately the significance of success or failure in one part is difficult to determine.

The Stanford-Binet tests were all given by the writer in the manner described by Terman. It is unnecessary to repeat this test in order to establish its reliability as the reliability has been independently reported upon by Terman.

The vocabulary and memory span for digits of the Stanford-Binet were given with the Porteus tests, the remainder of the Stanford-Binet taking only one session.

The Alpha 2 Reading Scale was scored by the method worked out by Kelley and his tables were used.

The tapping test was scored for number of taps and errors.

In the construction tests number of moves and time were taken and when the test was not completed within the limit of five minutes it was scored as a failure and the number of moves up to that time was noted.

A construction test—once solved—is much easier to solve a second time unless the first solution was due to chance. Healy A was repeated in order to check the first performance. In Healy's construction test B a second trial generally brings a result as near perfect as possible (that is, dependent only on skill and speed in motor performances), even if the first solution was hit upon by chance. It is impossible to do away with the chance element in performance tests, but in order to guard against it as much as possible, two tests were used each time, and the selection was made after a study of many types. There are several difficulties in making this choice and we were impressed by the fact that most performance tests have not been standardized and that there are very few tests of this kind which are sufficiently difficult for older subjects. The Healy and Knox tests satisfied both of these conditions.

The scoring for the learning tests is rather complicated. A perfect score on all four tests is four hundred, one hundred being the perfect score for each test. Learning test A has twelve elements, and if these were all correct on the three trials, thirty-six elements would receive a mark of one hundred, or each would get 2.8. Thus the score equals the number correct multiplied by 2.8. When a perfect score is made on the first or second trial, it is assumed that further trials would give a perfect score also. In learning test B there are five

symbols in each of the three trials,—consequently each receives a value of 6.7. In test C there are seven symbols and three trials. Dividing one hundred by three times seven there results a value of 4.7 for each, while in test D, which has ten items, the total number is thirty, with a value of 3.3 each. The total for all the tests is the sum of the score on each of the four.

RESULTS

Where a clinician is generally satisfied to take the score obtained by applying a test as a final goal, if in fact he goes so far as to work out a score, it is obvious that to attain the purposes here in mind the scores of the various tests used must be compared to gather statistics reflecting their qualities. That is, when the one hundred and sixteen children had been given the tests that were selected and when the scores were recorded, the field work was completed, but there remained to investigate in a laboratory manner what a combination of the results would show with reference to the purposes of this study.

This comparison of results was made by correlation, that is, by measuring the mutual implications (see Thorndike, Mental and Social Measurements, pp. 156-185). A test is to be evaluated in three ways; its correlation with criteria other than results of tests; its self-correlation, and its correlation with other tests. In the present inquiry we obtained no outside criteria with which to correlate our tests, because no outside criteria available could be relied upon. In the field of mental abilities, the only criteria which have been widely used are teachers' opinions, school marks, etc. These are unsatisfactory at best. Although we possess all these data for our cases we consider them useless since the children attended eight different schools in four places, with the marking systems varying for each. We compared judgments as to intelligence made by the teacher of the ninth grade of the Woodmere school with those made by the eighth grade teacher of the New York Public School. In the former the I. Q.'s varied from 95 to 141; in the latter from 73 to 116. In the former all but two children tested as supernormal and the class average was 121, whereas in the latter only one tested above 110 with a class average of 96. But to read the teachers' judgments one would think that the pupils of the latter school were considerably more intelligent than those of the former. Even the comparative ratings within one group were markedly unreliable. They showed all the errors of judgment pointed out by Terman. No account was taken of age; the best behaved, most conscientious pupil was invariably considered the most intelligent, etc. What is the use of making correlations with this kind of material,

when one knows in advance that all the fault of a low correlation will be attributed to the criterion, and the tests will stand as before—unknown quantities! Moreover, these criteria could only be used to represent a measure of general intelligence. The teachers admittedly knew practically nothing about the special abilities of their pupils; the parents, where consulted, knew very little more. A rating on general intelligence has been frequently correlated with general intelligence tests, and the results published. Our data would present no new factors.

Consequently we have not evaluated the tests by means of correlation with outside criteria but we do have the data for self-correlations and for inter-correlations. Where various tests which we used intercorrelate extremely highly, we may feel that they are measuring the same thing. On the other hand, if the intercorrelations approach zero or are negative, the results indicate that we have no evidence that aspects of intelligence are being measured at all. Only if the correlations are sufficiently high to indicate that intelligence is being measured and low enough to show that different factors are entering into the different tests, can we consider the tests worthy of being included in mental examination. In judging our correlations we must remember that we are testing normal children only,—therefore our coefficients are lowered—and that our ages do not cover a large area, which also lowers the coefficients of correlation.

Our conclusions are limited to the tests we used but the general method of dealing with the scores has a wide applicability.

Table IV

DISTRIBUTION OF PINTNER SCORES—100 CASES

Score	Frequency	Score	Frequency
200–209.9	0	370–379.9	3
210–219.9	0	380–389.9	5
220–229.9	1	390–399.9	2
230–239.9	0	400–409.9	4
240–249.9	0	410–419.9	5
250–259.9	1	420–429.9	5
260–269.9	0	430–439.9	7
270–279.9	1	440–449.9	4
280–289.9	1	450–459.9	5
290–299.9	2	460–469.9	7
300–309.9	1	470–479.9	5
310–319.9	1	480–489.9	1
320–329.9	2	490–499.9	4
330–339.9	4	500–509.9	6
340–349.9	2	510–519.9	4

TABLE IV—CONTINUED

Score	Frequency	Score	Frequency
350–359.9	6	520–529.9	1
360–369.9	8	530–539.9	2

16 were not given the Pintner Test.
The evenness of distribution of scores is noticeable.
Average=420.964. Unreliability 6.9.
Mean Square Deviation=68.95. Unreliability 4.9.

Table V

DISTRIBUTION OF MYERS SCORES—90 CASES

Score	Frequency	Score	Frequency	Score	Frequency
16	0	46	0	76	1
17	1	47	1	77	1
18	0	48	3	78	0
19	0	49	1	79	0
20	1	50	2	80	0
21	0	51	1	81	2
22	1	52	4	82	1
23	1	53	4	83	0
24	0	54	2	84	1
25	1	55	2	85	0
26	1	56	0	86	0
27	1	57	3	87	2
28	2	58	2	88	0
29	2	59	2	89	0
30	0	60	2	90	0
31	1	61	1	91	1
32	0	62	3	92	0
33	2	63	3	93	0
34	2	64	3	94	1
35	0	65	2	95	0
36	3	66	0	96	0
37	1	67	0	97	0
38	1	68	2	98	0
39	0	69	3	99	0
40	1	70	1	100	0
41	1	71	1	101	0
42	0	72	1	102	0
43	4	73	1	103	1
44	0	74	0	104	0
45	2	75	0	105	0

26 were not given test.
Average=53.325 Unreliability 1.8.
Mean Square Deviation=17.88. Unreliability 1.3.

Table VI

DISTRIBUTION OF ALPHA SCORES—107 CASES

Score	Frequency	Score	Frequency	Score	Frequency
3.6	2	5.4	1	7.3	7
3.7	0	5.5	1	7.4	6
3.8	0	5.6	1	7.5	12
3.9	0	5.7	1	7.6	2
4.0	0	5.8	0	7.7	6
4.1	3	5.9	2	7.8	1
4.2	1	6.0	0	7.9	2
4.3	0	6.15	1	8.0	2

Table VI—Continued

Score	Frequency	Score	Frequency	Score	Frequency
4.4	0	6.2	3	8.1	1
4.55	1	6.3	0	8.2	2
4.6	0	6.4	2	8.3	2
4.7	3	6.5	1	8.4	1
4.8	1	6.6	3	8.5	2
4.9	2	6.7	5	8.6	0
5.0	1	6.8	4	8.7	0
5.1	3	6.9	3	8.8	1
5.2	6	7.0	1	8.9	0
5.3	0	7.1	0	9.0	1
		7.2	7		

9 were not given The Alpha Test.
Average=6.834. Unreliability .116.
Mean Square Deviation=1.20. Unreliability .082.

TABLE VII

DISTRIBUTION OF PICTORIAL COMPLETION TEST SCORES—110 CASES

Score	Frequency	Score	Frequency	Score	Frequency
-15 to 0	2	30 to 34.99	6	65 to 69.99	12
0 to +5	2	35 to 39.99	6	70 to 74.99	7
5 to 9.99	2	40 to 44.99	6	75 to 79.99	4
10 to 14.99	1	45 to 49.99	10	80 to 84.99	9
15 to 19.99	4	50 to 54.99	8	85 to 89.99	4
20 to 24.99	4	55 to 59.99	12	90 to 94.99	2
25 to 29.99	0	60 to 64.99	8	95 to 99.99	1

6 were not given test.
Average=54.527. Unreliability 2.16
Mean Square Deviation=22.69. Unreliability 1.5.

TABLE VIII

LEARNING TESTS DISTRIBUTION—106 CASES

Score	Frequency	Score	Frequency	Score	Frequency
170–179.9	2	250–259.9	4	330–339.9	4
180–189.9	0	260–269.9	7	340–349.9	4
190–199.9	1	270–279.9	6	350–359.9	9
200–209.9	0	280–289.9	5	360–369.9	7
210–219.9	2	290–299.9	8	370–379.9	9
220–229.9	1	300–309.9	10	380–389.9	2
230–239.9	4	310–319.9	8	390–399.9	4
240–249.9	1	320–329.9	7	400–409.9	1

10 were not given the tests; 3 none at all; 7 not all four.
Average=300.66. Unreliability=5.08.
Mean Square Deviation=52.31. Unreliability=3.6.

TABLE IX

PORTEUS SCORES DISTRIBUTION—113 CASES

Score	Frequency	Score	Frequency	Score	Frequency
5	2	8.5	4	11.5	14
5.5	1	9	2	12	7
6	0	9.5	6	12.5	15
6.5	1	10	6	13	15
7	3	10.5	8	13.5	4
7.5	4	11	11	14	7
8	3				

3 were not give this test.
Average=11.09. Unreliability .19.
Mean Square Deviation=2.02. Unreliability .13.

TABLE X

DISTRIBUTION OF CROSSLINE TEST SCORES—114 CASES

Score I II	Frequency	Score I II	Frequency	Score I II	Frequency
Both OK^1	70	OK^2–OK^3	1	OK^1–F	3
OK^1–OK^2	13	OK^3–OK^2	2	OK^2–F	1
OK^2–OK^1	5	OK^3–OK^3	1	OK^3–F	2
OK^2–OK^2	5	OK^1–OK^4	2	OK^4–F	1
OK^1–OK^3	2	OK^2–OK^4	1	F –F	5

OK^1=Correct on first trial.
OK^2=Correct on second trial.
F=Failure on fourth trial.

TABLE XI

DISTRIBUTION OF TAPPING SCORES. AVERAGE OF 2 TRIALS—113 CASES

Score	Frequency	Score	Frequency	Score	Frequency
40 to 44.99	1	65 to 69.99	11	90 to 94.99	5
45 to 49.99	3	70 to 74.99	13	95 to 99.99	5
50 to 54.99	5	75 to 79.99	23	100 to 104.99	2
55 to 59.99	9	80 to 84.99	20	105 to 109.99	0
60 to 64.99	9	85 to 89.99	6	110 to 114.99	1

Average=73.43. Unreliability 1.26.
Mean Square Deviation=13.39 Unreliability .89.

TABLE XII

DISTRIBUTION OF CONSTRUCTION AND KNOX—TIME 108 CASES

Score	Frequency	Score	Frequency	Score	Frequency
50 to 99.99	1	350 to 399.99	12	650 to 699.99	6
100 to 149.99	7	400 to 449.99	8	700 to 749.99	4
150 to 199.99	5	450 to 499.99	8	750 to 799.99	2
200 to 249.99	10	500 to 549.99	9	800 to 849.99	3
250 to 299.99	10	550 to 599.99	5	850 to 899.99	2
300 to 349.99	8	600 to 649.99	7	900 to 949.99	0
				950 to 999.99	1

Average=420 to 480 or 7.685.
Mean Square Deviation=3.39.

Tables 4 to 12 inclusive show the distribution of scores on the various tests. The average, or more properly speaking the arithmetic mean and mean square deviation, are also given for each.

That we have sufficient cases is shown by the relation of the variability to the average. In only a few instances is it large enough to raise a doubt as to whether enough cases were used. These are the Pictorial Completion test, the Construction tests, and the Myers Mental Measure. The formula for the unreliability of an average is $\sigma T\text{–obt.av.} = \frac{\sigma dis.}{\sqrt{n}}$ for the unreliability of a mean square deviation it is $\sigma T\text{–obt}\sigma = \frac{\sigma dis.}{\sqrt{2n}}$ These data are also included in the tables. (See Thorndike, Mental and Social Measurements).

A few special considerations arose at once with reference to the crossline tests, the tapping test and the construction tests.

The crossline test has no value for our subjects (see table X.); one hundred and fourteen cases were tested, out of which 70, or over 60 per cent, made perfect scores; the remaining 40 per cent ranging almost indifferently from one error to complete failure. This test is, then, far too easy for our subjects, and the results are useless for our purposes. We will disregard it completely from now on.

In dealing with the tapping test we were confronted with the problem of how to handle the errors. Since a perfect correlation would be expected between two absolutely perfect tests of tapping ability, the highest correlation obtainable is presumably the one which best accounts for the errors. On this assumption the two trials of fifty cases of the tapping test were correlated both by Pearson and Spearman formulae, first disregarding the errors, then weighting them one each, and finally weighting them two points each, with the following results:

	Pearson	Spearman
Erors disregarded	r = .794	r = .917
Errors weighted one each	r = .773	r = .90
Errors weighted two each	r = .764	r = .82

It would seem then that the errors are of comparatively little importance, but as disregarding them gives the highest self-correlation, they will be omitted in any correlations in which the tapping test is involved.

A similar problem is presented by the construction tests, where we have scores for time and moves: Should they be combined and if so, how? If not, are they both important, or only one, and if the latter, which one? In order to arrive at an unbiased conclusion—for it was the writer's opinion that time was by far the most valuable measure—the advice of fifteen other persons was sought. These others were all familiar with the tests, and had used them extensively in clinical work. By far the majority were in favor of using both time and moves, each independently of the other. Two of these considered the moves decidedly more important than time; two others stated that time alone was sufficient, because time and moves had been found to correlate so highly, that the difference between using them and not doing so was within

the probable error of either one. None recommended attempting to combine them.

The following correlations were therefore made:

Construction A with B–time.
Construction A withB–moves.
Knox A with B–time.
Knox A with B–moves.
Average Construction A and B with average Knox A and B–time.
Average Construction A and B with average Knox A and B–moves.

If the test was not completed in five minutes it was scored as a failure and the number of moves up to that time recorded. Some children who solve the test in three minutes make more moves than others who fail in five minutes. How can one tell how many moves the latter would have made, had they completed the test? Obviously, the number they made until they were arbitrarily stopped is not a fair measure. It was finally decided to omit all cases where any construction test was a failure, from the moves correlations.

The crude scores were not used in the time correlations, but the three hundred seconds were divided into twenty groups of fifteen seconds each. Anyone succeeding with a test in fifteen seconds or less, was put in group one; if he took more than fifteen seconds and less than thirty-one seconds he was put in group two, and so forth. All who failed the test were put in group twenty, thus making it possible to include in these correlations many cases which had to be excluded from the correlations of number of moves made.

Taking up first the self-correlations, that is, the correlations of our alternate tests, with each other or the correlations of the scores obtained by repeated use of the same test, the results were as follows:

As only a few Stanford-Binet's were repeated, the results are of little significance. We obtain a correlation of .89 on our fourteen cases. L. M. Terman ("The Intelligence of School Children" ch. IX.) had retests given to three hundred and fifteen children, out of which forty-six were given three or more tests. The interval between the first and second testing ranged from one day to seven years. The central tendency of change is represented by an increase of 1.7 in

I.Q.; the middle fifty per cent of change lies between the limits of 3.3 decrease and 5.7 increase. Consequently the probable error of a prediction based on the first test is 4.5 points in terms of I.Q. The correlation between all the testings is .933. Apparently whether the interval be a few months or several years does not influence the result. If the re-examination be within a few days, the I. Q. will—on the average—be raised only two or three points, and this when no restriction has been put on the children communicating with one another. There are several exceptions to this general rule, one being that young feeble-minded children tend to show their feeble-mindedness more as they grow older; that is, they test lower on the Stanford-Binet. We need not concern ourselves with this, as only normal children were included in this study. Another obvious factor which tends to make the I. Q. appear unstable, is due to the fact that the test is limited at the upper end. As a child with a high I. Q. grows older, the I. Q. drops until at the age of sixteen years the highest I. Q. obtainable is 122. In many pathological cases such as children suffering from epilepsy, chorea, etc., the I. Q. fluctuates considerably. But even within the ranges of normality, Terman thinks that fluctuations occur for at least three reasons.

1. There may be a certain amount of irregularity in the actual rate of mental development.

2. The results of a test may be influenced to some extent by the conditions under which it is given, the state of the child's health, his attitude toward the test, fatigue and other temporary and accidental features.

Retests after a brief interval indicate that errors from this source are ordinarily not large.

3. There is inevitably a certain amount of error in every I. Q. rating due to imperfections in the scale used.

What has been generally criticized in the Stanford-Binet scale, namely that it measures different things at different years and consequently that a subject might do very well when his memory ability for example was tested, and very poorly when his reasoning ability came into the foreground a couple of years later, does not seem to be valid on actual findings. The theoretical argument against such a criticism is that so many age levels are tested each time that a subject will win and lose points in every branch which the test includes.

The Pintner and Myers tests were chosen to measure the same thing, and so we expected to find a high correlation between them. The Pearson coefficient of .584 was so unexpected that we felt that further investigation was needed. A closer study of the tests revealed the fact that their likeness rested on negative similarity; neither involved the use of language, but in other respects they apparently required different abilities. The Pintner test appeared more limited, more mathematical, involving concrete situations rather than generalizations while the Myers on the other hand was more general, but rather sketchy. In order to test the truth of this hypothesis, the six Pintner tests were intercorrelated and also the four Myers tests—see table. The average of the Pintner intercorrelations was .234, of the four Myers tests correlated each with all the others, .445. It will therefore be seen that the above explanation is unsatisfactory.

TABLE XIII

PINTNER TESTS INTERCORRELATED

	1	2	3	4	5	6	Composite
1		−.009	.392	.325	.183	.337	.618
2	−.009		.470	.022	−.022	.107	.396
3	.392	.470		.361	.224	.316	.757
4	.325	.022	.361		.035	.456	.625
5	.183	−.022	.224	.035		.307	.540
6	.337	.107	.316	.456	.307		.670
Average	.249	.126	.353	.240	.154	.305	
Composite	.618	.396	.757	.625	.540	.670	

Average of all above correlations, regarding signs + .234.
Average of all above correlations, without regarding signs + .238.
Probable Error of each correlation, approximately .05.
Number of cases 100.

The fact that correlations between the separate Tests are low, while those of each Test with the composite of all 6, are high, indicates merit in the Test as a whole.

TABLE XIV

MYERS MENTAL MEASURE INTERCORRELATIONS

	1	2	3	4	Composite	Average of all above correlations— + .445
1		.470	.469	.564	.786	Number of cases—89
2	.470		.346	.424	.796	
3	.469	.346		.403	.477	
4	.564	.424	.403		.775	
Composite	.786	.796	.477	.775		

The comment made concerning the previous Table—Pintner Tests—applies to some extent to the Myers Test also. However, the correlations between the separate tests are much higher than those found between the Pintner Tests.

For if the Pintner tests were all of the same nature, including the same factors, their intercorrelations would be high. On the other hand, if the Myers tests were general, their inter-correlations would be lowered. Just the opposite occurs; the Pintner intercorrelations are lower than the Myers. These correlations can probably be explained on another basis. In the Pintner series certain tests are easier than others, most esspecially the second and fourth, which lowers the intercorrelations. In the Myers Mental Measure all the tests with the exception of the third are of about the same difficulty, the grading being within the test, and this raises the correlations. It also seems probable that while the Pintner tests do measure more limited factors, each test may measure a different one, the type of material alone remaining the same.

On *a priori* grounds something of this sort seems likely, for the material is practically the same, the correlations are low, so the factors measured must be different.

It is true that in the Myers Mental Measure the ability to respond to the spoken word (directions) is part of the test, and it is possible that this is a special ability—calling forth something akin to the abilities necessary for success with the Stanford-Binet, even where the language itself is easily understood. Such a factor our data are unable to measure, but it is interesting in this connection to compare the correlation of the Stanford-Binet with Pintner and of the Stanford-Binet with Myers.

In devising the Pintner non-language test, the effort was made to have it extend from the lowest to the highest grades. This meant introducing tests such as the second, which is far too easy for a child after he has reached the fourth or fifth grade, and also others which were almost incomprehensible to the young child, as tests four and six. Since our subjects are for the most part past the fourth or fifth school grade, we would expect to find some sign of their maturity in the correlations. Reference to the table shows that test two correlates lower with all the other tests than any other single test. The one exception is the correlation of tests two and three, which—it will be remembered—are identical in form, the latter being different from the former only in degree of difficulty. Test six, on the other hand, correlates higher with every other test, than test two. This is as it should be: had we tested younger children the table would

probably have shown entirely different results. Incidentally, these findings show the importance of bearing in mind the nature of the group that is being studied when interpreting correlations. Each part was also correlated with the total test score, with high results throughout, with the exception of test 2. In looking at this table, one must feel that the test is a good one, for the intercorrelations of the separate tests are low, but with the composite they are high.

Before leaving the Pintner test, mention should be made of a study by Jeanette Chase Reamer, in which she retested over four hundred children with this test with slightly less than a two-year interval, and found a correlation of .726 between the relative positions which they occupied at each testing. The closeness of this correlation was a complete surprise to both her and to Professor Pintner.

With regard to the Myers Mental Measure intercorrelations, we find them all fairly high and regular. The most surprising thing about them is that tests three and four which appear far more similar than any other two tests in this series, should have one of the lowest correlations,—lower than four with one or four with two. Also, we see no reason why one and four should correlate higher than any other two. If the language factor were significant, we should find one and three (where audible directions must be followed for each separate unit of the test) correlating highly, and also two and four (where after the original directions the subject is left to himself). But as a matter of fact, one and two, and one and four are higher than one and three, and two and four. However the degree of difference between the various correlations is so small that these comparisons must be taken in a negative rather than a positive sense; that is, we might have expected the correlations to prove *something*, instead of which they prove nothing! With the composite the separate tests correlate very highly, as would be expected since the composite includes always the test being correlated with it, thus giving a perfect relation between two out of the five factors. Test three proves an exception here also, and we feel that the fault is the same as with Pintner two: it is too easy for our subjects.

The Porteus Maze Test when correlated with itself gives a correlation of .95, which is high and satisfactory.

The intercorrelations of the construction tests gave the

most disconcerting results of all. They seem to prove Professor Thorndike's assertion that no matter how many construction tests are used, one cannot do away with the chance element. If four construction tests, when correlated for time and moves, give only .16 for the former and .08 for the latter, it seems like a hopeless task to give sufficient tests to raise the correlation to the high 70's or 80's. This is indeed a problem, for the construction test as such is undoubtedly desirable.

Perhaps more important than these low numerical results, is the fact that combining the individual tests does not seem to operate to raise the correlations. Thus the two Healy tests when correlated for time, give a result of .21, the two Knox tests similarly correlated give .27, but the average of Healy tests with average of Knox tests shows a correlation of only .16. In attempting to explain these findings, it must be remembered that the Healy tests were given on one occasion, and the Knox tests at least one week later, both A and B on the same day.

If our results were due primarily to lack of reliability of the construction tests from day to day then scores from two construction tests given on the same day ought to show higher correlations than we found. If the lack of reliability of construction tests from day to day is not to be considered because of the generally low correlations, and so if it makes no difference whether all four tests are given on the same day or different days, then our average correlations should not turn out to be lower than the correlations of tests given on the same day because an increase in the number of factors generally operates to raise the correlations. Other correlations between various combinations of the construction tests were made and are recorded in table 16, but the results are no more enlightening than the ones we have discussed here.

We are assuming here that the solution of each of the four construction tests involves the same abilities, not that they are of equal difficulty. We have no evidence to prove that this is the case, but we do not see how any construction tests could be devised which, though different, were apparently more similar than these. However, Healy test A and Knox test B are more similar than any other combination. Knox's test was modelled directly from Healy's and is supposed to be more difficult. A correlation between these two gives us

minus .055 for time and .126 for moves. In other words, the correlation between the two is about what one would obtain between two factors having no relationship to each other at all. If this is true of Healy Construction test A and the Knox diamond-shaped frame test, we conclude that construction tests have no constant value for intelligence testing.

It will be recalled that the Thorndike Reading Scale, Alpha 2, was not repeated as the alternate scales now available had not yet been published, but we might quote Dr. McCall's statement to the effect that a high correlation was obtained between our scale and the more recently devised alternate on representative subjects.

No correlation was obtained between two trials of the Healy Pictorial Completion test II because the number of cases who were retested was small, about twenty-five, no further retesting being done because the attitude of the subjects was so different on the second testing that the repetition was more a matter of memory than anything else.

In repeating this test a week or more after the first presentation it was found that the correct pieces were again put in. Of those that were incorrect about half were the same, the other half being pictures having about the same value in scoring, so that the total score was very little altered. It was generally slightly increased, rarely lowered. The only other test of this kind available is Healy's Pictorial Completion test A, which is so simple that almost all of our subjects would make perfect scores on it. The attitude of the subjects, when this test was offered a second time, was not good. The test appeals because it is a new situation presenting a problem in an attractive form. The second time, the newness has worn off. The usual response is, "I've done that before," or words to that effect. If the child is urged to attempt a better performance, he will often ask in a surprised tone of voice, "Didn't I do it perfectly before?" Even when one succeeds in getting a child to try again, he rarely makes any effort, but puts in at once the pieces selected before or similar ones. If he comments audibly on his performance, it runs something like this, "Oh, that one,—a book was missing there; where is it? Here—why there are two—well it doesn't matter, it's a book he dropped." Occasionally one will notice that it *does* matter, but even this is due largely to chance, to his happening to have spied two books this time.

As a whole it was felt that what was gained by repeating this test in the way of establishing its reliability, was not equivalent to what was lost in the attitude of the subjects to the tests as a whole. If repeated at the very end of the testing, this difficulty would be in part eliminated, but it was decided to omit its repetition completely.

The reliability of the Healy and Bronner learning tests was not ascertainable as no alternate series has been devised, and as no other test could be found which appeared sufficiently similar to warrant the hypothesis that it measured the same thing.

The tapping test was repeated in exactly the same form, and showed an intercorrelation of .81 with a P. E. of .022. This we may consider a satisfactory correlation, showing that the test has a high degree of reliability.

The self-correlations having been thus completed and analyzed, the next step is to consider the intercorrelations of the tests. Let us now consider the correlation of each test with the Stanford-Binet mental age. As all the tests are given crude scores regardless of age, in order to have comparable data, the mental age must be used instead of the I. Q.

TABLE XV

	Stanford-Binet	(probable error)	Number of Cases
Pintner	.439	±.055	97
Myers	.686	±.037	89
Alpha 2	.757	±.027	106
P. C. 11	.541	±.045	110
Learning	491	±.049	105
Porteus	.536	±.045	110
Tapping	.604	±.040	112
4 Construction	.426 (Time)	±.078	107
4 Construction	.326 (moves)	±.097	83
Healy A	.410 (Time)	±.079	110
Healy A	.374 (moves)	±.092	88
Healy B	.281 (Time)	±.088	110
Healy B	.088 (moves)	±.105	88
Knox A	.046 (Time)	±.095	111
Knox A	.009 (moves)	±.099	103
Knox B	.216 (Time)	±.090	111
Knox B	.112 (moves)	±.098	103

Total number of tests correlated with Stanford-Binet r = .5976. Woodworth's method of combining the results of several tests used, Av. $r = \frac{M\ Av.S^2-1}{m-1}$ (Woodworth: Combining the Results of Several Tests).

The first column stands for the Pearson coefficient obtained

from the formula $r = \frac{\Sigma (x.y)}{N\sigma_A \sigma_B}$ or, as it is usually stated, $r = \frac{\Sigma (x.y)}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}}$. The P. E. in this case means the probable divergence of the true coefficient of correlation from that obtained from a limited random selection of cases. The formula was was σT–obt. $r = \frac{1-r^2}{\sqrt{n}}$. If the median deviation of the probable divergence is desired it may be obtained by multiplying the figures in the second column by.6754. For a discussion of these formulae and any other statistical methods here used, see E. L. Thorndike, Mental and Social Measurments.

It is interesting to find the Alpha test correlating most closely with the Standford-Binet of all the tests used. It corroborates to some extent the current opinion that the Stanford-Binet is largely a test of language ability.

The next highest correlation, that of the Myers Mental Measure, is more difficult to explain. Although intended as a non-language group intelligence test, it involved more language than any of the other tests employed. Still it would seem surprising if this were such a tremendous factor. It tends to indicate the validity of group tests as does also Alpha, in that these two tests were given to nearly all of the subjects in groups, and yet correlate more highly with the individual Stanford-Binet than any of the other tests do, practically all of which were given individually.

One of the most surprising results is the high correlation of tapping with the Stanford-Binet. One would generally assume that the type of motor ability required in our tapping test had little to do with intelligence—especially with older subjects. Our data apparently contradict this hypothesis, and we are confronted with the necessity of explaining the data. It is known that tapping ability increases with chronological age at least up to maturity in the absence of tremors, epilepsy, chorea and other diseases affecting the co-ordinating mechanisms. When we consider that our subjects were all normal and therefore their mental ages tended to increase with their chronological ages, and that all our cases were treated together regardless of age, it at once seems plausible that we have here a spurious correlation due to increase in both scores with chronological age, rather than intelligence.

We have therefore correlated tapping with the Stanford-Binet I. Q.'s, which represent intelligence regardless of age, the coefficient obtained being .069, and find that our assumption is justified.

The correlations between the construction tests and Stanford-Binet are very low, the only reasonably high coefficient being obtained with Healy A. This correlation was about the same as the composite of construction tests with Stanford-Binet. It is interesting to note that Healy A was the only construction test which Professor Terman used in his revision inasmuch as he considered that one only to meet the requirements sufficiently to be included.

The remaining tests, Pintner, Porteus, Learning, and P. C. 2, that have been correlated with Stanford-Binet, each show a correlation very close to .50. This we consider significant in that they are sufficiently high to show that we are measuring intelligence, restricting that term to its generally used meaning with reference to mental testing. In addition the coefficients of correlation are low enough so that we may conclude that different abilities of the subjects tested are being measured, that is, the use of different tests does not result in a repeated measurement of the same abilities. Consequently the use of these tests in addition to the Stanford-Binet, means the measurement of more varieties of ability than can be tested by the Stanford-Binet alone. It remains to be determined whether Pintner, Porteus, Learning, and P. C. 2 all measure the same factor or whether some if not all of them can be used to distinguish special abilities which the others do not test. The answer to this inquiry lies in the results obtainable by the correlation of all of these tests with each other. These results are recorded in Table 16 and they are results so unexpected that they call for interpretation.

Many of the correlations recorded in the table show that the importance of the language factor has been overestimated in dealing with older school children. The correlation between language and non-language tests are high enough to show that the language factor need not be avoided to have a test which can be said to measure intelligence.

Let us first consider the correlation between Myers and Alpha 2; the former is supposed to be a non-language test, the latter a test of understanding of sentences. If language were an important factor it would be hard to account for a

correlation of .733, the second highest obtained aside from the self-correlations. P. C. II is a performance test dealing with pictures; it is concrete where the Alpha 2 deals with abstract ideas,—yet these two give a correlation of .709—likewise unquestionably high. We have stated elsewhere that some language enters into the Myers Mental Measure, and that ability to respond to the spoken word *may* be an exceedingly important factor. If so, why does P. C. II give a coefficient of .714 with Myers Mental Measure? If the Myers Mental Measure is a non-language test, why is the correlation of Pintner, a thorough going non-language test, with Myers lower than Pintner with Alpha 2, a test involving so much language? Again, when we compare the correlation of Alpha 2 and Porteus—.701—with that of P. C. II and Porteus—.702—we are at a loss to explain the similarity in result unless we discard the idea of the importance of the language factor. For the Porteus test requires no language.

One must not overlook the importance of language as a handicap in giving tests to foreigners, etc., but where older school children are being tested it cannot be vital. For in order to succeed in the higher grammar school grades, it is essential that they have a fairly good working knowledge of the English language, and this is all that is needed to succeed with the so-called language tests.

Coming to the selection of a schedule of tests we conclude that:

1. Reading scale Alpha 2 should be included. In the first place because of its high correlations with other tests, the highest of any test with all the others, and also because of special considerations. It must be remembered that Alpha 2 is entirely a reading and writing test and therefore one would not expect so uniformly high a correlation as exists between it and the other tests which are supposed each to be specially adapted toward bringing out certain abilities. The high correlations remind us of Binet's constant contention that intelligence, broadly speaking, can be tested by language tests. This conclusion, however, does not imply that a non-language test cannot serve a like purpose. Our intercorrelations show that there was no reason to avoid language tests inasmuch as they correlated highly with the non-language tests. This is interesting in that the tendency in devising tests is towards making them language tests. For causes of this tendency

we can ascribe first, the simplicity and lack of apparatus inherent in them and second, that the difficulty or ease of the test is far more readily regulated than in the non-language tests. In scrutinizing a test to forecast the results of its use our results seem to show that it is not necessary to dwell upon whether or not the tests involve the use of language.

2. The list of tests selected includes both Myers and Pintner. Myers correlates more highly with every other test than does Pintner, with the exception of their respective correlations with Porteus. As has been stated, both of these tests are valuable and in addition they have the merit of yielding different results.

3. The intercorrelations of P. C. II with such of the other tests as we found to be reliable were sufficiently high to make us believe that this test should be included in our schedule in spite of the uncertainty as to whether it is reliable.

4. Definite judgment upon the learning tests should be reserved for the present. Their highest correlation is under 50 and we have no evidence that they are reliable. Before the learning tests have a right to be so called they must be shown really to measure learning ability; they must also be tested for reliability. It is a question whether learning ability of a given individual is uniform in all fields. The uniformity of learning ability cannot be assumed, for a mere assumption as to the uniformity of motor ability proved to be wrong (See Perrin, An Experimental Study of Motor Ability). If learning ability is found not to be so, combining the various tests may operate to conceal what is valuable in them.

5. The tapping test should be included, in the discretion of the examiner, not because the results can be relied upon to indicate intelligence but because giving this test, which takes only a minute, may disclose latent defects in motor control.

In an attempt to reach some definite conclusion about the construction tests we have made many correlations of different combinations. Healy A is the only test which gives a correlation as high as .40 and that with the Stanford-Binet. Nor does combining the tests raise the correlations, for the four construction tests correlated with Stanford-Binet give a result practically no higher than Healy A alone. It seemed useless to correlate the construction tests with the other tests when they gave such unsatisfactory results with each other, and with the Stanford-Binet. We have no evidence that these four

tests are valuable either as intelligence tests, or for any other purpose.

The Porteus test is one of the most interesting. Since no language enters into the test, one would expect it to correlate more highly with Pintner and Myers than with the Stanford-Binet and Alpha 2. Just the opposite occurs; of the four, by far the highest correlation is with Alpha 2. The correlation of Porteus and P. C. II is practically the same. These three tests all seem to call for one kind of ability. Is it good judgment, common sense ability, planfulness, deliberation, carefulness, foresight, good apperceptions? Probably it contains these and other similar traits. It is the difference between these tests which brings the correlations down to .70, and which causes them to correlate differently with the other tests. Alpha 2 and Stanford-Binet, both requiring language, correlate more highly than Porteus and Stanford-Binet, or P. C. II and Stanford-Binet. Some other factor causes Alpha 2 and P. C. II to correlate considerably higher with Myers than Porteus does. There are always many traits measured by every test, no matter how simple, and the emphasis on the different factors is not always the same for the same test. It varies with the group being measured; their age, sex, education, social selection, etc. Why does Porteus get a higher correlation between his test and Binet's than we do? Partly, at least, because he tested children of all ages, but especially younger ones, whereas ours group themselves closely about a mode, and are older.

We have quite a number of cases which are not completely measured by either the Stanford-Binet or the Porteus tests; that is, they could probably succeed with some harder tests if they were given the opportunity, and this lowers our correlations. The fact that many of Porteus' cases were placed higher rather than lower on his tests than on the Stanford-Binet seems to show that they tended to be poorer in language ability than in planfulness, apperceptions,—whatever one wishes to call it. Our cases on the other hand seem to find no difficulty with the language factor. It is by comparing the results of the same group in different tests, and of different groups on the same test, that most can be learned of what the tests actually do measure. In this study we have the same group measured by many different tests. We find our intercorrelations high in many cases, but nowhere so high that we

feel that the tests are identical. However certain similarities such as the one just discussed between the Porteus, P. C. II, and Alpha 2, were brought to light by this method. Differences such as the striking one between Pintner and Myers have also been observed. If different groups had been used, one would be unable to draw any conclusions regarding the tests for the groups themselves might be responsible for so many of the factors. Again different factors of the tests are brought out by different groups as for instance a younger and older set of children tested with the Pintner non-language survey test would give an entirely different kind of intercorrelation between the separate tests. By this method we can take account of more factors and so interpret our findings with greater accuracy.

There is no evidence that the P. C. test measures apperceptions, that the learning tests measure learning ability, that the construction tests measure ability to use concrete material. On this account, and also because each involves too many incidental, disturbing factors, none of these tests can be considered adequate measures of special abilities or disabilities. Such tests are much needed, and should be constructed so as to measure fundamental, underlying differences in ability. They must be correlated with everything of any possible importance in order to ascertain the degree to which one ability is related to all others. In studying memory we want to know how important a part it plays in reasoning, in mechanical work, etc. We must learn the significance of a good memory for every school study, and for various occupations. If different kinds of memory play important parts in different studies and vocations, this too we must find out. It is a big task, perhaps impossible to carry out at present, but without such information we are tremendously handicapped. The taboo of "faculty" psychology has contributed to lessen activity along these lines, for if you investigate memory you are getting perilously near something obsolete. But very few would deny that there is such a thing as remembering, and all study of memory and its ramifications has yielded interesting and important results.

It has been more or less tacitly assumed in the past that differences in performance are due to differences in the material used rather than to underlying "faculty" differences. This was based upon findings such as those obtained when

TABLE XVI

INTERCORRELATIONS

	ST-B	Pintner	Myers	Alpha 2	P.C.II	Porteus	Learning	Tapping
ST-B		.439	.686	.757	.541	.536	.491	.604
Pintner	.439		.584	.597	.423	.356	.268	.282
Myers	.686	.584		.733	.714	.330	.321	.522
Alpha 2	.757	.597	.733		.709	.701	.420	.332
P. C. II	.541	.423	.714	.709		.702	.290	.382
Porteus	.536	.356	.330	.701	.702	.95	.457	.483
Learning	.491	.268	.321	.420	.290	.457		.252
Tapping	.604	.282	.522	.332	.382	.483	.252	.805

ST.-B. I. Q. with Tapping r = .069.
ST.-B. M. A. with Total number of remaining Tests r = .5976.
ST.-B. with average Pintner and Myers r = .588.
Healy A with Knox A—Time—r = .135.
Healy A with Knox B—Time—r = -.055.
Healy A with Knox B—moves—r = .126.
Knox A with Knox B—Time—r = .27.
Knox A with Knox B—moves—r = .103.
Healy A with Healy B—Time—r = .21.
Healy A with Healy B—moves—r = -.001.
Average Healy A and B with average Knox A and B—Time—r = .161.
Average Healy A and B with average Knox A and B—moves—r = .076.

This Table summarizes the results of the correlations of each Test with every other. It should be noted that a correlation was obtained between Stanford-Binet and all other Tests combined the coefficient of .5976 being fairly high.

memory was tested. It was found that a good memory for logical material did not follow from a good memory for nonsense; that being able to remember visually presented facts did not necessarily indicate ability to remember what was heard. The result of these and similar observations has been the development of tests dealing with specific types of material, or—giving up the specific side entirely—tests of general intelligence. Our data seem to indicate that real, underlying differences do exist, if we only know how to get at them. In order to prove this, it is necessary to have a test with omnibus material, all of which is designed to measure a certain type of thing. We shall now proceed to do this.

A COMBINATION TEST FOR PLANFULNESS

The correlations in table XVI, particularly those obtained between Porteus, Alpha 2, and P. C. II, seem to indicate the possibility of a factor, common to all and largely determining the score on each, which has nothing to do with the material employed, that is, whether a language or non-language test, or the like. We have suggested above several names for this factor,—good judgment, common sense, deliberation, carefulness, foresight, good apperceptions, planfulness, persistence, prudence and mental alertness in meeting a new situation, ability to see the whole of a situation instead of reacting to the most obvious part of it. An attempt was made to investigate it more thoroughly by combining the elements of each test which seemed most specifically to measure it. The selection was made from the Porteus, Myers, Alpha 2, P. C. II, and Stanford-Binet tests. All the tests selected would require about twenty-five minutes to perform, this being a liberal estimate based upon the time limit for each test. Alpha 2 has no definite time limit, but from the writer's experience, ten minutes would seem ample to allow for the parts of the test included in this selection. When all the individual tests had been chosen, they were divided into two sections, and a self-correlation of .763 was obtained with 80 cases. The tests in each group were:

I. Porteus—year 11 (scored 0, 1, 2) year 12 (scored 0, 1, 2, 3, 4).
Myers—pages 4. Numbers 3 and 7 (scored each 0, 1).
P. C. II—pictures 2 and 6 (scored 1 each if OK; otherwise 0).
Alpha 2, Part II—difficulty 8—number 4 (scored 0, 1).
Pintner—test 5, numbers 5 and 7 (scored each 0, 1).
Pintner—test 6, picture 2 pieces 2 and 1 (scored each 0, 1).
Pintner—test 6, picture 3, pieces 4 and 1 (scored each 0, 1).

II. Porteus—year 10 (scored 0, 1, 2) year 14 (scored 0, 1, 2, 3, 4).

Myers—page 4. Numbers 5 and 10 (scored each 0, 1).
P. C. II—pictures 7 and 8 (scored 1 each if OK; otherwise 0).
Alpha 2, Part II—difficulty 8—number 1 (scored 0, 1).
Pintner—test 5, number 6 (scored 0, 1).
Pintner—test 6, picture 2, pieces 4 and 3 (scored each 0, 1).
Pintner—test 6, picture 3, pieces 2 and 3 (scored each 0, 1).
Stanford-Binet—XIV years, number 6 (scored 0, 1).

The Porteus tests were chosen because they were devised to measure this very thing. The fact that only one type of material—mazes—was included, was considered by Porteus one of the outstanding advantages of his test. We feel that this is a disadvantage since some children might have a disability for working with this kind of material although possessed of common sense, foresight, etc. With omnibus material this special factor is overcome. The choice of the four most difficut tests was largely a matter of the distribution of the subjects. Too many would have made perfect records on the easier tests.

The selection from Myers Mental Measure was based largely upon resistance to suggestion. In each case four pictures with some element in common must be chosen from eight possible ones and underlined. These four could not be too difficult or our subjects would all score 0; if they were too easy we would have no reason to believe that this characteristic pertained to them. Number 3 is the selection of four toys, —a tricycle, top, kite and rocking horse, with a soldier as the confusing picture. In number 5, four items made of iron must be chosen,—a stove, dagger, or sword, train, and lock. This has several confusing suggestions. There is a broom which might be associated with the stove, and two animals which might be connected with the train as they all are capable of locomotion. Number 7 consists of an insect, a broom, a bird, a table, a butterfly, an aeroplane, a goat and a cow. The four things which can travel in air are to be underlined. The two animals prove confusing to many children. In number 10 the subject is to select four articles of wood,—two trees, a barrel and a table, with a snake, a camel, a cannon, and a

bird to be omitted. Here also the three animals receive considerable attention, the hasty child not noticing that the fourth is lacking, or the snake is overlooked, the two remaining animals and two trees being classed together as objects possessing life. There seems to be some suggestion in each of these pictures, and it is certainly true that a careful, deliberate, performance by a subject who takes in the whole situation and responds to it will give far better results than a hasty, careless one.

The pictures from P. C. II are those in which there are several obvious possibilities. A hasty, careless selection will hit upon the first possible one, rather than searching further for the exactly correct one. All correct pieces were checked by asking the subject why that particular one had been chosen and if it was put in by chance, no credit was given. The partial credits given by Healy were omitted, the picture scored either as perfect or a failure. This was necessary in order to eliminate the other possible factors which enter into solving the test partly. For instance, in the second picture, where a book is missing, it is not sufficient to put in any book, pencil case or lunch box, but by following up persistently all the clues, the one and only correct red book can be placed in the space with certainty.

From the Thorndike Alpha 2 reading scale questions were selected which had been answered by a large number of children. Question I requires a fairly careful study of the paragraph in order to find just what it is that seems true at first but is really false. The question is a little clumsily put,—certainly not direct and to the point,—which is an advantage for our purposes. Question 4 is not a reading scale problem proper, but necessitates close attention to several directions. In two rows of digits the subject must underline every five that comes just after a two, unless the two comes just after a nine. If that is the case, he must draw a line under the next figure after the five. The last few lines of the first page of the Myers Mental Measure are similar to this, but the Alpha 2 was given to a larger number of cases, there was no time limit, and less possibility of copying, so it was given the preference, as being more accurate.

Numbers 5, 6, and 7 from Pintner test 5 are all similar in nature. Given a drawing, the problem is to draw it in a reversed position, with two lines of the second position given

on which to construct the rest. This seems like a rather special ability, but Pintner gives each drawing considerable weight in his total score, and persistence and planfulness are certainly essential for a good performance.

Pintner test 6 consists of parts of pictures presented in a disarranged order. Each part is numbered and blank spaces are provided in which the subject is to place the numbers of the parts in order which would give a perfect ensemble. Here again planfuness, patience, and foresight are needed, and on the whole the subject who possesses them to the greatest degree will be the most successful.

Finally one test was selected from the Stanford-Binet scale, —namely the reversed clock hands of year XIV. If two out of three were correct a score of one was given, if less no credit at all. This test seemed to require the same kind of ability as many of the other tests included, and was therefore added. Some of the other Stanford-Binet series might have been used also, but those which seemed desirable came too high or too low in the scale so that the distribution for our subjects would not be satisfactory.

The correlation of .763 obtained between the two parts is fairly high when it is remembered that the highest score on each section can only be 17; also that the whole series of both parts would only take half an hour to give. As to reliability it is a noteworthy conclusion that this self-correlation is the highest one obtained with any non-identical material. A correlation of the composite tests with any of the tests which are included would probably give a high coefficient difficult to interpret because of the varying amount of each included in the composites, and a low correlation with learning tests, construction tests, or tapping could hardly be considered strong evidence in favor of our new grouping. But the correlation with Stanford-Binet seemed worth finding, and when worked out yielded a coefficient of .537. This indicates that our combination test is comparable with the whole series of tests from which it was compiled. We have, however, no criterion to prove that it actually measures the trait which we presuppose it does. But this same criticism applies to all the tests which are supposed to measure specific factors. Our new test combination of old material is certainly as good as the tests from which it originated; we think it is better, because it gives evidence of measuring one trait, or group of

traits with a variety of materials, whereas all the others measure many kinds of traits with identical or similar material. That is, the classification and material preparatory to the formation of a test has generally heretofore been along the lines of the material employed, such as form boards, etc., whereas the combination test being discussed presents the results obtained from forming a test directed toward planfulness, or other ability.

CONCLUSION

It is proposed to set forth the practical results of this study, to show the positive information that has been ascertained and also to show from the experience gathered in the course of obtaining such information, what further investigations should be made, with what purpose, and what methods may lead to success. This study has reached some positive results and has disclosed other perhaps more valuable ones in the same field.

In entering upon this study it was believed that the results of the method that has been pursued would justify the conclusion that the Stanford-Binet series can be used as a test of general intelligence and that certain other tests used as auxiliaries would make apparent and give a measure of special abilities not individually measured by the Stanford-Binet. It was expected that the various tests would give reasonably high correlations with the Stanford-Binet and rather low correlations with each other, thus on the one hand establishing the reliability of the tests used, and on the other hand, the diversity of the abilities that were subjected to measurement.

These results were anticipated because care was used in selecting the tests to take those which had an approved authorship, an extended use, a definite purpose, and a general reputation of success in the field they purported to cover. That is, the various units had each been shown apparently to be satisfactory and on these *a priori* grounds it was thought that properly selected units used in conjunction would result in a reliable schedule.

Had the results of the correlations been in harmony with this anticipated situation, we might properly have pointed to this study as a demonstration of the process by which schedules of tests for children should be composed.

Looking upon our results as they have been reported upon, the fact is obvious that there is no such easy manner in which to arrive at reliable schedules of tests. Unexpected low correlations were obtained in some situations where the indicated results should have been high, and vice versa, and while our positive purpose therefore met with disappointing obstacles, a study of the figures as we have them led to other worthwhile conclusions.

Drawing upon the results of the correlations, it can be stated with assurance that it will not be well to take tests upon which a high face value has been placed when they were used without being effectively valued by comparison, and combining a number of them in the expectation of using the combination to get reliable information as to the general intelligence and the special abilities of normal children. One of the best examples which we can show, as a result of this study, of the impropriety of such procedure is, that the type of material used does not govern the abilities tested. We obtained a higher correlation between a language and a non-language test than between two language tests or two non-language tests, similar examples can be drawn from the correlations listed above respecting other characteristics of various tests. Insofar therefore as authors of tests have relied upon the material as a quality that would single out and measure a certain one of many abilities, it seems clear that individual tests miss their purpose. However, the correlations did seem to show that something definite was being tested, so that if our purpose of finding a schedule of tests at once sufficient to measure both general and special abilities, was disappointed, at least the schedule we used can be relied upon for general abilities and that such a schedule is more reliable than the Stanford-Binet alone. The components of this schedule have been previously listed and it only remains to state what individual matters of interest relating to each were made clear in the course of the study which was directed to larger purposes.

It was a matter of actual demonstration herein that all of the construction tests used are unreliable, this conclusion disproving the previously held opinion based upon empirical considerations to the effect that they reliably measure ability to handle concrete material.

Persons having occasion to apply mental tests have too frequently overlooked the matter of how far the test can be relied upon. This is an important matter and consequently it should be of some interest to note that the reliability of the Stanford-Binet, Pintner non-language group test, Thorndike reading scale Alpha 2, Porteus Maze Test, and tapping test has been established, whereas the Myers Mental Measure, the Healy Pictorial Completion test II, the Healy-Bronner

learning tests and the crossline tests are not yet definitely shown to be reliable.

Care should also be observed in interpreting the results of correlations, for the mere fact of high correlation is only generally and not conclusively proof of reliability. There is the possibility that factors causing unreliability have been hidden—thus, in the tapping test, the high correlation with Stanford-Binet was deceptive owing to the fact that the scores on both increased with the age of the subjects. Other specific remarks relating to individual tests are contained in the results.

There remains to state what considerations we have found to have a probable value as to future work in this field. If we found on the one hand that the type of material used in a test does not govern the ability tested, on the other hand there are some indications that to test individual abilities the test should have a variety of material. So far the elements of a desired test can be stated, but the further necessity of finding just what material is suitable, can only be determined by practical work consisting of correlation with outside criteria and with any other measures of claimed effectiveness in the field in question.

As an experimental example, for the confines of this study would allow no more extended investigation, various parts of a number of the tests were united in a combination test intended to secure a measure of planfulness. The resulting correlations indicated success in this attempt. A similar or even greater measure of success may follow further combinations aimed at the measurement of other abilities.

It may also be stated as having been illustrated in the course of this study that the supposed merit of various mental tests based upon various, insufficient or unscientific criteria, such as mere hypothesis, or even practical results, if relied upon, may lead to misleading or dangerous conclusions, and that before one takes the responsibility of giving advice or of taking action with respect to information gained from the application of mental tests, there should be available the assurance that proper comparative tests and correlations have verified the supposed propriety of relying upon the results.

VITA

The author of this dissertation was born in New York, August 25, 1898. Secondary education was at Far Rockaway High School, taking highest honors, and receiving Regents Scholarship for College. Vassar College, 1915-1917; Barnard College, Columbia University, 1917-1919; B. A. Degree Columbia University, 1919, Honors in Psychology; 1918, research work for New Jersey State Institution for Feeble Minded; 1919-1920, Fellowship at Judge Baker Foundation, Boston, Assistant Psychologist; Columbia University, 1920-1922, Post Graduate Work in Psychology.